PROJECT REPORT ON:

"Micro-Credit Defaulter Model"

SUBMITTED BY:

Shweta Kumari

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my SME Shwetank Mishra
as well as Flip Robo Technologies who gave me the opportunity to do this project on Rating Prediction, which also helped me in doing lots of research wherein I came to know about so many new things especially the data collection part.

Also, I have utilized a few external resources that helped me to complete this project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project .

Introduction:

Business Problem Framing:

- A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.
- Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.
- Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.
- We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

Conceptual Background of the Domain Problem

Telecom Industries understand the importance of communication and how it affects a person's life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour.

They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah).

The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

We have to build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label '1' indicates that the loan has been payed i.e. Non- defaulter, while, Label '0' indicates that the loan has not been payed i.e. defaulter.

Review of Literature

An attempt has been made in this report to review the available literature in the area of microfinance. Approaches to microfinance, issues related to measuring social impact versus profitability of MFIs, issue of sustainability, variables impacting sustainability, affect of regulations of profitability and impact assessment of MFIs have been summarized in the below report. We hope that the below report of literature will provide a platform for further research and help the industry to combine theory and practice to take microfinance forward and contribute to alleviating the poor from poverty.

Motivation for the Problem Undertaken

I have to model the micro credit defaulters with the available independent variables. This model will then be used by the management to understand how the customer is considered as defaulter or non-defaulter based on the independent variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand whether the customer will be paying back the loaned amount within 5 days of insurance of loan. The **relationship between predicting defaulter and the economy** is an important motivating factor for predicting micro credit defaulter model.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

In this particular problem I had label as my target column and it was having two classes Label '1' indicates that the loan has been payed i.e, Non- defaulter, while, Label '0' indicates that the loan has not been payed i.e. defaulter. So clearly it is a binary classification problem and I have to use all classification algorithms while building the model. There was no null values in the dataset. also, I observed some unnecessary entries in some of the columns like in some columns I found more than 90% zero values so I decided to drop those columns. If I keep those columns as it is, it will create high skewness in the model. To get better insight on the features I have used ploting like distribution plot. With this ploting I was able to understand the relation between the features in better manner. Also, I found outliers and skewness in the dataset so I removed outliers using percentile method and I removed skewness using yeo-johnson method. I have used all the classification algorithms while building model then tunned the best model and saved the best model. At last I have predicted the label using saved model.

Data Sources and their formats

The sample data is provided to us from our client database. It is hereby given to us for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

Also, my dataset was having 209593 rows and 36 columns including target. In this particular datasets I have object, float and integer types of data. The information a bout features is as follows.

Features Information:

1.label: Flag indicating whether the user paid back the credit amount within 5 day s of issuing the loan{1:success, 0:failure}
2.msisdn: mobile number of user
3. aon : age on cellular network in days
4. daily_decr30 : Daily amount spent from main account, averaged over last 30 da ys (in Indonesian Rupiah)
5. daily_decr90 : Daily amount spent from main account, averaged over last 90 da ys (in Indonesian Rupiah)
6. rental30 : Average main account balance over last 30 days
7. rental90 : Average main account balance over last 90 days
8. last_rech_date_ma : Number of days till last recharge of main account
9. last_rech_date_da: Number of days till last recharge of data account
10. last_rech_amt_ma : Amount of last recharge of main account (in Indonesian R upiah)
11. cnt_ma_rech30 : Number of times main account got recharged in last 30 days
12. fr_ma_rech30 : Frequency of main account recharged in last 30 days
13. sumamnt_ma_rech30 : Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)
14. medianamnt_ma_rech30 : Median of amount of recharges done in main accou nt over last 30 days at user level (in Indonesian Rupiah)
15. medianmarechprebal30 : Median of main account balance just before recharg e in last 30 days at user level (in Indonesian Rupiah)
16. cnt_ma_rech90 : Number of times main account got recharged in last 90 days

17. fr_ma_rech90 : Frequency of main account recharged in last 90 days
18. sumamnt_ma_rech90 : Total amount of recharge in main account over last 90 days (in Indonasian Rupiah)
19. medianamnt_ma_rech90 : Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah)
20. medianmarechprebal90 : Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah)
21. cnt_da_rech30 : Number of times data account got recharged in last 30 days
22. fr_da_rech30: Frequency of data account recharged in last 30 days
23. cnt_da_rech90 : Number of times data account got recharged in last 90 days
24. fr_da_rech90 : Frequency of data account recharged in last 90 days
25. cnt_loans30 : Number of loans taken by user in last 30 days
26. amnt_loans30: Total amount of loans taken by user in last 30 days
27. maxamnt_loans30 : maximum amount of loan taken by the user in last 30 days
28. medianamnt_loans30 : Median of amounts of loan taken by the user in last 30 days
29. cnt_loans90 : Number of loans taken by user in last 90 days
30. amnt_loans90 : Total amount of loans taken by user in last 90 days
31. maxamnt_loans90 : maximum amount of loan taken by the user in last 90 days
32. medianamnt_loans90 : Median of amounts of loan taken by the user in last 90 days
33. payback30 : Average payback time in days over last 30 days
34. payback90 : Average payback time in days over last 90 days
35. pcircle : telecom circle
36. pdate : date

Data Preprocessing Done

- As a first step I have imported required libraries and I have imported the dataset which was in csv format.
- Then I did all the statistical analysis like checking shape, nunique, value counts, info etc….
- Then while looking into the value counts I found some columns with more than 90% zero values this creates skewness in the model and there are chances of getting model bias so I have dropped those columns with more than 90% zero values.
- While checking for null values I found no null values in the dataset.
- I have also dropped Unnamed:0, msisdn and pcircle column as I found they are useless.
- Next as a part of feature extraction I converted the pdate column to pyear, pmonth and pday. Thinking that this data will help us more than pdate.
- In some columns I found negative values which were unrealistic so I have converted those negative values to positive using abs command.
- Also, I have converted all the float values in maxamnt_loans90 to zero as it is specified in the problem statement we can have only 0,6,12 as maximum amount of loan taken by the user in last 30 days. As well I have dropped all the data with amnt_loans90=0 as it gives the persons who have not taken any loans.

Data Inputs- Logic- Output Relationships

- Since I had all numerical columns I have plotted dist plot to see the distribution of each column data.
- In maximum features relation with target I observed Non-defaulter count is high compared to defaulters.

Hardware & Software Requirements and Tools Used

While taking up the project we should be familiar with the Hardware and software required for the successful completion of the project. Here we need the following hardware and software.

Hardware required: -

1. Processor — core i5 and above
2. RAM — 8 GB or above
3. SSD — 250GB or above

Software/s required: -

1.Anaconda

Libraries required:-

To run the program and to build the model we need some basic libraries

In [1]:
```
#importing required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt

import warnings
warnings.filterwarnings('ignore')
```

As follows:

- **import pandas as pd**: **pandas** is a popular Python-based data analysis toolkit which can be imported using import pandas as pd. It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a numpy matrix array. This makes pandas a trusted ally in data science and machine learning.
- **import numpy as np**: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
- **import seaborn as sns:** Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.
- **Import matplotlib.pyplot as plt:** matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
- from sklearn.linear_model import LogisticRegression
- from sklearn.preprocessing import StandardScaler
- from sklearn.ensemble import RandomForestClassifier
- from sklearn.tree import DecisionTreeClassifier
- from xgboost import XGBClassifier
- from sklearn.ensemble import GradientBoostingClassifier

Data Analysis and Visualization

Identification of possible problem-solving approaches (methods)

- To remove outliers I have used percentile method. And to remove skewness I have used yeo-johnson method. We have dropped all the unnecessary columns in the dataset according to our understanding. Use of Pearson's correlation coefficient to check the correlation between dependent and independent features. Used VIF to check for Multicollinearity. Also I have used Normalization to scale the data. After scaling we have to balance the target column using oversampling. Then followed by model building with all Classification algorithms. I have used oversampling (SMOTE) to get rid of data imbalancing. The balanced output looks like this.

Testing of Identified Approaches (Algorithms)

- Since label was my target and it was a classification column with 0-defaulter and 1-Non-defaulter, so this particular problem was Classification problem. And I have used all Classification algorithms to build my model. By looking into the difference of accuracy score and cross validation score I found BaggingClassifier as a best model with least difference. Also to get the best model I had to run through multiple models and to avoid the confusion of overfitting I had go through cross validation. Below are the list of classification algorithms I have used in my project.
- XGBClassifier
- DecisionTreeClassifier
- BaggingClassifier
- AdaBoostClassifier
- Logistic Regression

Key Metrics for success in solving problem under consideration

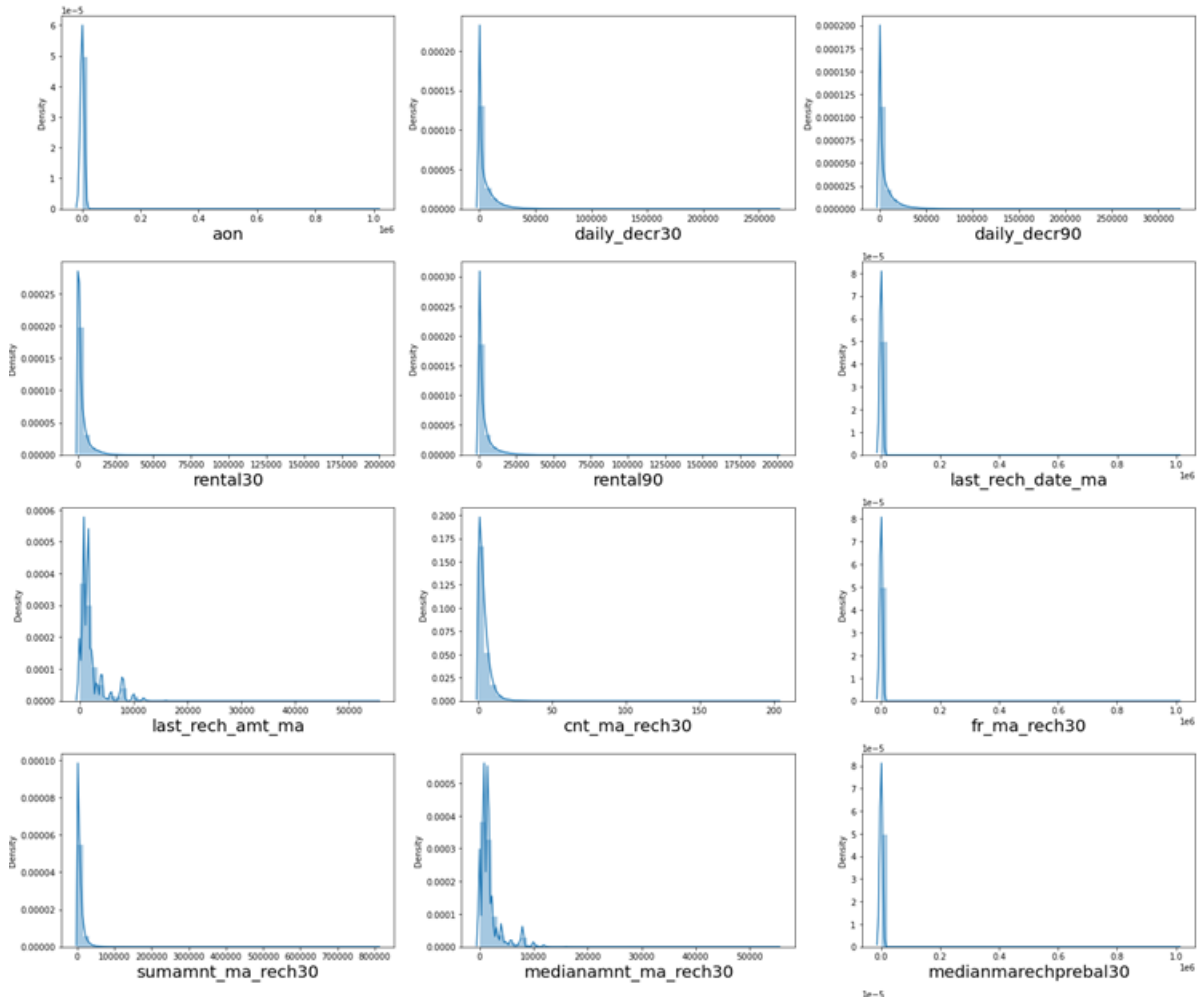
I have used the following metrics for evaluation:

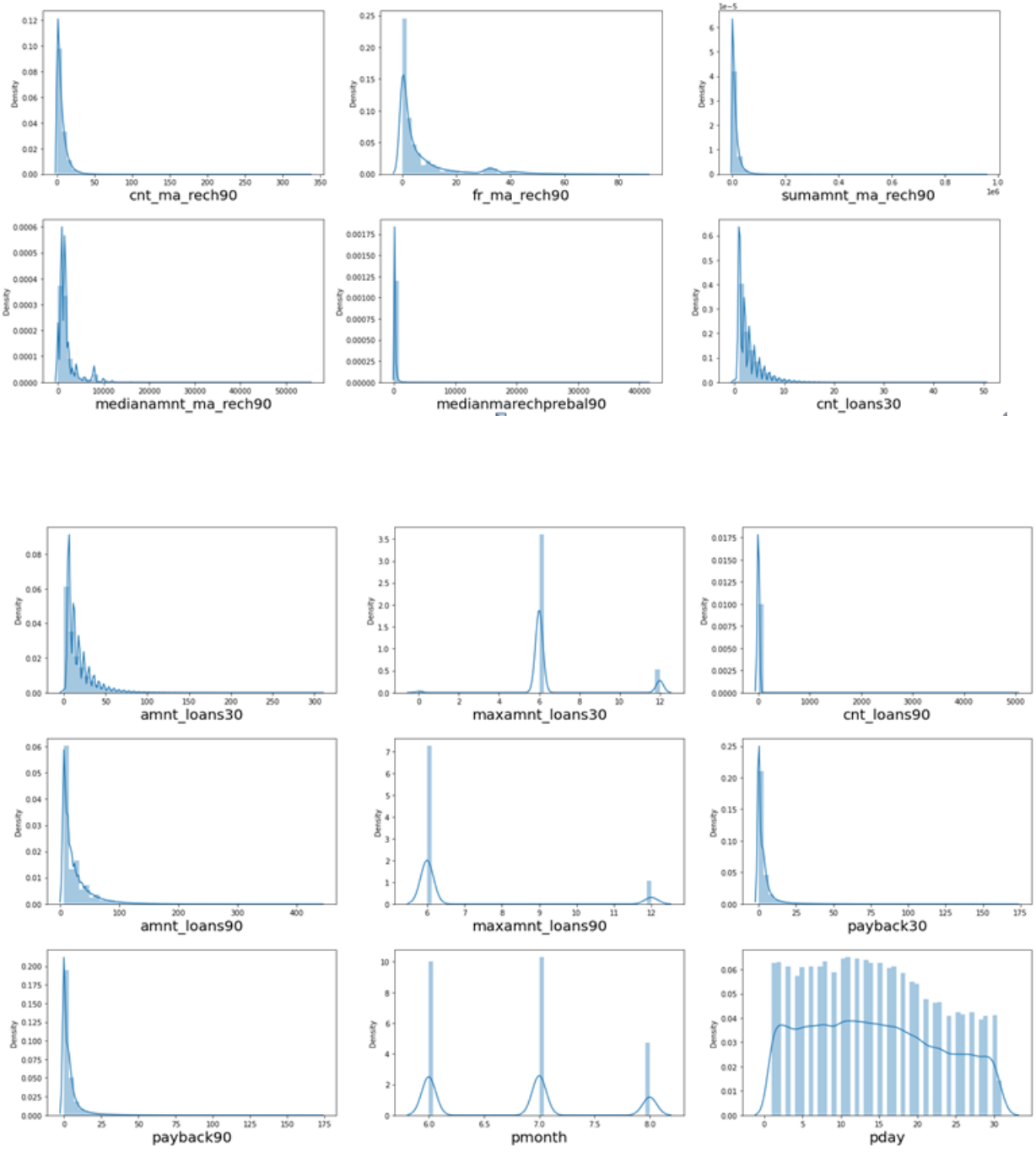
- **Precision** can be seen as a measure of quality, higher precision means that an algorithm returns more relevant results than irrelevant ones.
- **Recall** is used as a measure of quantity and high recall means that an algorithm returns most of the relevant results.
- **Accuracy score** is used when the True Positives and True negatives are more important. Accuracy can be used when the class distribution is similar.
- **F1-score** is used when the False Negatives and False Positives are crucial. While F1-score is a better metric when there are imbalanced classes.
- **Cross_val_score**: To run cross-validation on multiple metrics and also to return train scores, fit times and score times. Get predictions from each split of cross-validation for diagnostic purposes. Make a scorer from a performance metric or loss function.
- **AUC_ROC _score**: ROC curve. It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0
- I have used accuracy_score since I have balanced my data using oversampling.

Visualization

I have used bar plots to see the relation of numerical feature with target and I have used 2 types of plots for numerical columns one is disp plot for univariate and bar plot for bivariate analysis.

Univariate Analysis for numerical columns:





Observations:

- We can clearly see that there is skewness in most of the columns so we have to treat them using suitable method
- There is a data imbalancing issue so we have to treat this by using oversampling or under sampling.

Run and Evaluate selected models

1) XGB Classifier:

```
Accuracy Score: 94.54852847941342
Confusion Matrix: [[51004  3238]
 [ 2695 51896]]
              precision    recall  f1-score   support

           0       0.95      0.94      0.95     54242
           1       0.94      0.95      0.95     54591

    accuracy                           0.95    108833
   macro avg       0.95      0.95      0.95    108833
weighted avg       0.95      0.95      0.95    108833


Cross validation score : 93.4929894371647
\Accuracy_Score - Cross Validation Score : 1.055539042248725
```

XGBClassifier is giving me almost 94.5% accuracy.

2) DecisionTree Classifier:

```
Accuracy Score: 91.17730835316495
Confusion Matrix: [[49879  4363]
 [ 5239 49352]]
              precision    recall  f1-score   support

           0       0.90      0.92      0.91     54242
           1       0.92      0.90      0.91     54591

    accuracy                           0.91    108833
   macro avg       0.91      0.91      0.91    108833
weighted avg       0.91      0.91      0.91    108833


Cross validation score : 90.79571844280974
\Accuracy_Score - Cross Validation Score : 0.38158991035520273
```

DecisionTreeClassifier is giving me 91% accuracy.

3) BaggingClassifier:

```
Accuracy Score: 93.71606038609613
Confusion Matrix: [[51607  2635]
 [ 4204 50387]]
              precision    recall  f1-score   support

           0       0.92      0.95      0.94     54242
           1       0.95      0.92      0.94     54591

    accuracy                           0.94    108833
   macro avg       0.94      0.94      0.94    108833
weighted avg       0.94      0.94      0.94    108833


Cross validation score : 93.22945162071272
\Accuracy_Score - Cross Validation Score : 0.4866087653834086
```

BaggingClassifier is giving me almost 94% accuracy.

4) AdaBoostClassifier:

```
Accuracy Score: 84.92093390791396
Confusion Matrix: [[47032  7210]
 [ 9201 45390]]
              precision    recall  f1-score   support

           0       0.84      0.87      0.85     54242
           1       0.86      0.83      0.85     54591

    accuracy                           0.85    108833
   macro avg       0.85      0.85      0.85    108833
weighted avg       0.85      0.85      0.85    108833


Cross validation score : 84.92569522312496
\Accuracy_Score - Cross Validation Score : -0.004761315210998873
```

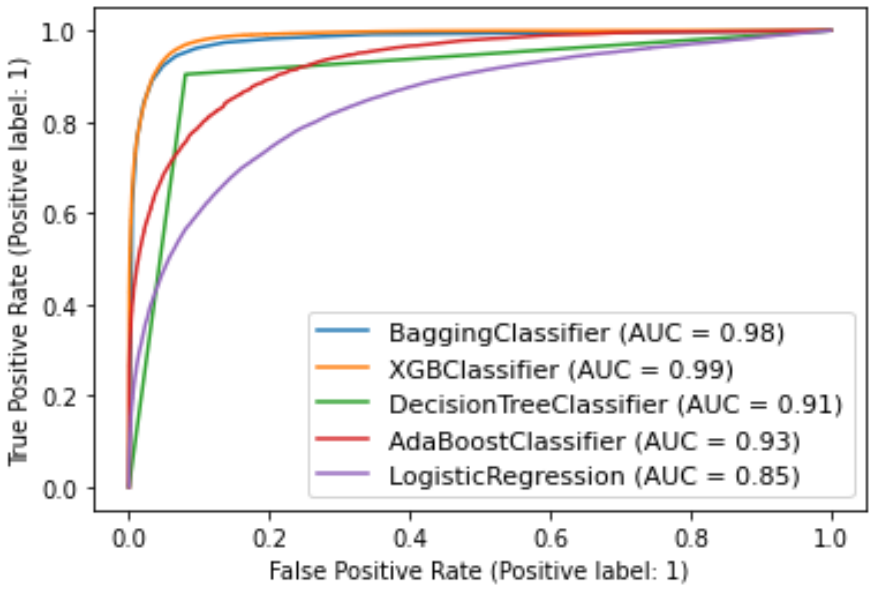
AdaBoost Classifier is giving me 85% accuracy.

5) Logistic Regression:

```
Accuracy Score: 77.10528975586449
Confusion Matrix: [[42778 11464]
 [13453 41138]]
              precision    recall  f1-score   support

           0       0.76      0.79      0.77     54242
           1       0.78      0.75      0.77     54591

    accuracy                           0.77    108833
   macro avg       0.77      0.77      0.77    108833
weighted avg       0.77      0.77      0.77    108833


Cross validation score : 77.07704060849807
\Accuracy_Score - Cross Validation Score : 0.0282491473664237
```

Logistic Regression is giving me 77% accuracy.

ROC-AUC Curve:



- AUC value is high for XGBClassifier and BaggingClassifier.I got least difference in model accuracy and cross validation score for BaggingClassifier so BC is my best model.

Hyper Parameter Tunning:

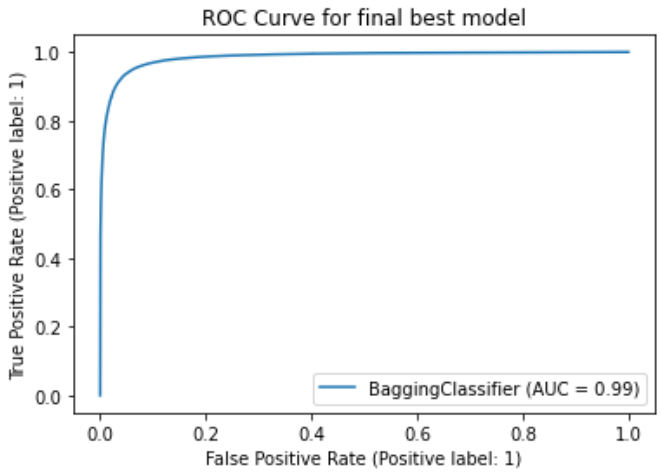
```
GCV=GridSearchCV(BaggingClassifier(),parameter,cv=5)
GCV.fit(X_train,y_train)

Final_mod=BaggingClassifier(bootstrap='True', n_jobs=-1,warm_start='True', n_estimators=40)
Final_mod.fit(X_train,y_train)
pred=Final_mod.predict(X_test)
acc=accuracy_score(y_test, pred)

print('Accuracy Score:',(accuracy_score(y_test,pred)*100))
print('Confusion matrix:',confusion_matrix(y_test,pred))
print(classification_report(y_test,pred))
```

```
Accuracy Score: 94.37211140003492
Confusion matrix: [[51399  2843]
 [ 3282 51309]]
              precision    recall  f1-score   support

           0       0.94      0.95      0.94     54242
           1       0.95      0.94      0.94     54591

    accuracy                           0.94    108833
   macro avg       0.94      0.94      0.94    108833
weighted avg       0.94      0.94      0.94    108833
```

Our model accuracy has increased from 93.7% to 94.37% that's good.

- I have choosed all parameters of BaggingClassifier, after tunning the model with best parameters I have incresed my model accuracy from 93.7% to 94.37%.

ROC Curve for final model:



- After hyperparameter tuning we got improvement in roc curve and AUC also.

CONCLUSION

Key Findings and Conclusions of the Study

In this project report, we have used machine learning algorithms to predict the micro credit defaulters. We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the features. Thus we can select the features which are correlated to each other and are independent in nature. These feature set were then given as an input to four algorithms and a hyper parameter tunning was done to the best model and the accuracy has been improved. Hence we calculated the performance of each model using different performance metrics and compared them based on these metrics. Then we have also saved the best model and predicted the label. It was good the the predicted and actual values were almost same.

Learning Outcomes of the Study in respect of Data Science

I found that the dataset was quite interesting to handle as it contains all types of data in it. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. New analytical techniques of machine learning can be used in property research. The power of visualization has helped us in understanding the data by graphical representation it has made me to understand what data is trying to say. Data cleaning is one of the most important steps to remove unrealistic values and zero values. This study is an exploratory attempt to use four machine learning algorithms in estimating micro credit defaulter, and then compare their results.

To conclude, the application of machine learning in micro credit is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to crediting institutes, and presenting

an alternative approach to the valuation of defaulters. Future direction of research may consider incorporating additional micro credit transaction data from a larger economical background with more features.

Limitations of this work and Scope for Future Work

- First draw back is the length of the dataset it is very huge and hard to handle.
- Followed by more number of outliers and skewness these two will reduce our model accuracy.
- Also, we have tried best to deal with outliers, skewness and zero values. So it looks quite good that we have achieved a accuracy of 94.82% even after dealing all these drawbacks.
- Also, this study will not cover all Classification algorithms instead, it is focused on the chosen algorithm, starting from the basic ensembling techniques to the advanced ones.